Internet2 QBone - Building a Testbed for Differentiated Services

Authors

Benjamin Teitelbaum (ben@internet2.edu)
Internet2 (UCAID) / Advanced Network & Services

Susan Hares (skh@merit.edu)
Merit Network

Larry Dunn (ldunn@cisco.com)
Cisco Systems

Vishy Narayan (vnarayan@mail.arc.nasa.gov)
NREN/NGI Program, Raytheon/NASA Ames Research Center

Robert Neilson (rneilson@bcit.bc.ca)
British Columbia Institute of Technology

Fracis Reichmeyer (freichme@nortelnetworks.com)
Nortel Networks

Table of Contents

Abstract

The Internet2 project is a partnership of over 130 U.S. universities, 40 corporations and 30 other

organizations. Since its inception, one of the primary technical objectives of Internet2 has been to engineer scalable, interoperable, and administrable interdomain Quality of Service (QoS) to support an evolving set of new advanced networked applications. Applications like distance learning, remote instrument access and control, advanced scientific visualization, and networked collaboratories will allow universities to fulfill their research and education missions into the future, but only if the network QoS that these applications require can be assured. To meet this challenge, the Internet2 QBone initiative [QBone] has brought together a dedicated group of U.S. university and federal agency networks, international research networks, engineers, researchers, and applications developers to build a testbed for interdomain IP Differentiated Services (DiffServ).

Requirements for Internet2 Quality of Service

Although the QBone initiative is still quite young, it rests on more than a year of collective head scratching by members of the Internet2 QoS Working Group. Starting in fall 1997, this working group (including more than 30 distinguished networking experts from academia, industry, and government) has struggled to understand the QoS requirements of advanced networks and how Internet2 could begin to make progress toward meeting them. At a series of workshops in fall 1997 and winter 1998, the working group heard from advanced applications developers, campus network planners, and gigaPoP operators, and identified a demanding set of requirements for Internet2 QoS. Chief among these are:

- **Relevance to Advanced Applications**
  End-to-end QoS services must meet the absolute performance requirements of advanced applications.

- **Scalability**
  Any viable QoS architecture must scale well both with respect to the large number of flows and high forwarding rates of core routers, as well as with respect to administrative burden.

- **Interoperability**
  Any viable QoS architecture must allow multiple, independently configured and administered implementations of services to be concatenated to form well-defined notions of end-to-end QoS. In particular, they must allow for interoperability among implementations provided by many different equipment vendors.

Differentiated Services

In parallel with the working group's efforts, the Differentiated Services approach to QoS began to attract significant interest from the IETF. Although much of the push for DiffServ at this time was from commercial ISPs who saw a sizable and immediate market for differentiated classes of best-effort IP service, architectural engineering concerns played a significant role as well. From a technical perspective, DiffServ is a reaction against the perceived scalability problems of the IETF Integrated Services (IntServ) model. DiffServ is an attempt to find simple, scalable forms of QoS that can provide a variety of end-to-end services across multiple, separately administered domains, without necessitating complex interprovider business arrangements or complex

behaviors in the forwarding equipment.

The DiffServ architectural framework achieves its scaling properties by indicating in each packet's header [RFC2474] one of a few standardized, simple differentiated forwarding treatments. These simple aggregate packet treatments, also known as per-hop behaviors (PHBs), are combined with a much larger number of policing policies enforced at the network edge to provide a broad and flexible range of services, without requiring state or complex forwarding decisions in core routers.

Each DiffServ micro-flow is policed and marked at the first trusted downstream router according to a contracted service level agreement (SLA), usually a token bucket filter. The QBone does not interfere with the strictly bilateral negotiations of SLAs. For this reason, the term "service level description" (SLD) is often used to refer to what the QBone itself defines. However, in order to emphasize the requirement that this kind of specification be part of all SLAs between QBone participants, we continue to use the term SLA in what follows.

When viewed from the perspective of a network administrator, the first trusted downstream router is a leaf router at the periphery of the trusted network. Downstream from the nearest leaf router, a DiffServ flow is mingled with similar DiffServ traffic into a behavior aggregate; all subsequent forwarding and policing is performed on aggregates. At inter-provider boundaries, service level agreements specify the transit service to be given to each aggregate. Aggregate SLAs are also characterized by traffic profiles (again, often based on token bucket filters). By carefully enforcing the aggregate traffic contracts between clouds and ensuring that new reservations do not exceed aggregate traffic capacity, the DiffServ architecture provides well-defined end-to-end services over concatenated chains of separately administered clouds. Furthermore, since SLAs exist only at the boundaries between clouds, the result is a set of simple bilateral service level agreements that mimics current interprovider exchange agreements.

In addition to packet forwarders capable of implementing the emerging PHB standards, the Differentiated Services architecture [RFC2475] requires edge devices that implement classifying, metering, marking, shaping, and dropping. Although not currently part of the DiffServ architecture, it is expected that a new kind of network component known as a bandwidth broker (BB) will play an important role in automating admission control for DiffServ networks.

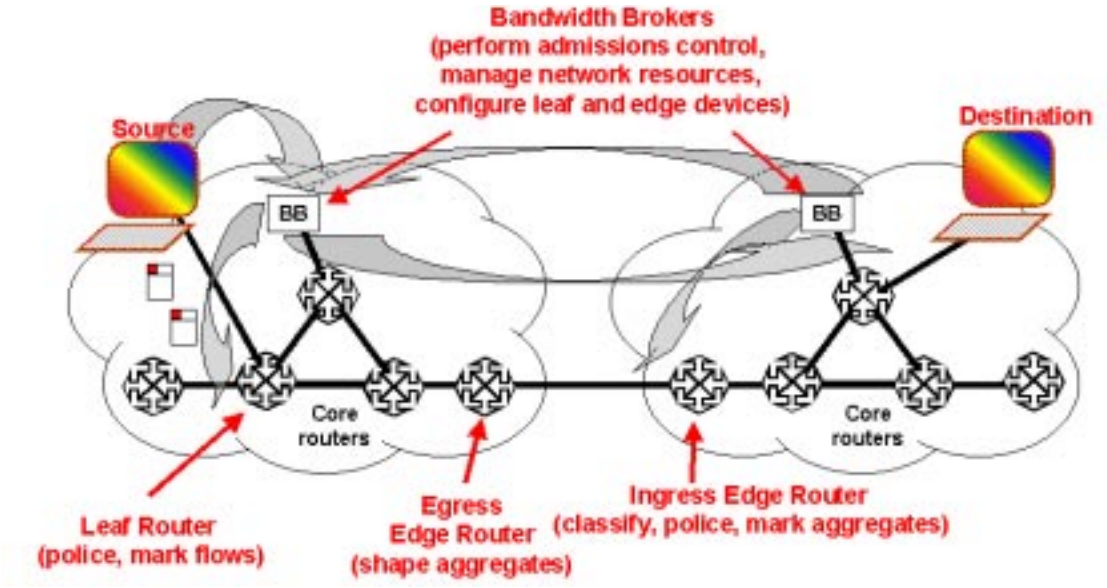

Figure 1. The Differentiated Services Architectural Framework

Per-flow policing and marking is performed by the first trusted leaf router downstream from the sending host. When a local admission control decision has been made by the sender's cloud, the leaf router is configured with the contracted per-flow service profile. Downstream from the first leaf router, all traffic is handled as aggregates. At cloud ingresses, incoming traffic is classified according to a Traffic Conditioning Agreement (TCA) into behavior aggregates, which are policed according to the SLA in place. Depending on the particular DiffServ service model in question, out-of-profile packets are either dropped at the edge or marked for a different PHB.

In order to reduce burstiness, it is very important that each cloud which initiates a QoS flow have the ability to do its own traffic shaping. In addition, as in-profile traffic traverses a cloud, it may experience induced burstiness caused by queuing effects or increased aggregation. Consequently, clouds may need to shape on egress to prevent otherwise conforming traffic from being unfairly policed at the next downstream cloud.

Finally, to make appropriate internal and external admission control decisions, and to configure leaf and edge device policers correctly, each cloud may be outfitted with a bandwidth broker. When a sender signals its local BB to initiate a reservation, the requesting user is authenticated and subject to a local admission control decision. On behalf of the sender, the BB may then initiate an end-to-end reservation request along the chain of bandwidth brokers representing the clouds to be traversed by the flow. The bandwidth broker abstraction is critical because it allows separately administered network clouds (possibly implemented with very different underlying Layer-2 technologies and subject to very different policies) to manage their network resources as they see fit.

Because QoS creates a valuable resource that must be protected against theft, security is an essential consideration. Luckily, the DiffServ architecture greatly simplifies the situation. The edge-core functionality split ensures that the policy enforcement points (PEPs) in any security model will only be the edge devices. Nodes in the interior of a DiffServ domain are responsible only for per-hop differentiated forwarding based on the DiffServ codepoint (DSCP) and not for

enforcing service level descriptions. Furthermore, because inter-domain reservations are built out of concatenations of bilateral reservations for aggregates, the number of trust relationships in DiffServ is relatively small. A more complete discussion of security issues appears in the discussion of bandwidth brokers below.

Recognizing the startling similarity between the root engineering motivations behind DiffServ and the primary Internet2 QoS requirements, the Internet2 QoS Working Group began to develop an Internet2 QoS architecture based on the evolving IETF DiffServ architecture. In May 1998, at the First Internet2 Joint Applications/Engineering QoS Workshop [I2QoS98], the working group presented the outlines of this architecture and began a dialogue within the greater Internet2 community. This dialogue culminated in consensus around the need to build an interdomain testbed to explore and advance DiffServ technologies and to iteratively provide an increasingly robust infrastructure for experimentation with new advanced collaborative applications.

Options for Achieving End-to-End Resource Allocation

Within the context of the evolving DiffServ architecture there is a spectrum of proposed mechanisms for effecting end-to-end resource allocation. This section points out the tradeoffs involved with the various methods under consideration. Subsequent sections provide detail on the particular choices that will be evaluated in the initial QBone.

Each of the methods below selects a particular balance among: which device does packet marking, how much signaling is involved (either host-BB, BB-BB, or BB-router), the expected frequency of signaling, and the degree to which resource allocation for a flow or aggregate of flows is recognized end-to-end. The methods below are roughly ordered from those offering weak assurances in exchange for minimal treatment, to those offering strong assurances but that require more involved treatment.

The spectrum of mechanisms includes:

1. Do nothing: This is just current best effort (BE) delivery. It is mentioned to establish a pole in the spread of possible options. No marking, no signaling, no local or end-to-end resource allocation.

2. Layer-2 treatment locally, static inter-AS bandwidth allocation: Again, not quite DiffServ, this idea uses IEEE 802.1p treatment in the campus network to give packets better treatment via Layer-2 marking. No explicit DS-byte marking is done, no dynamic signaling, some local resource allocation. Inter-AS links are monitored, and expanded as necessary to give adequate performance. Such a method is discussed further by Terry Gray in [I2QoS98].

3. Host DS-byte marking, no signaling: This is a minimalist DiffServ approach; while requiring that a host mark packets, the rest of the provisioning is straightforward. Layer-3 devices might be configured in a variety of static ways: from a single DSCP always being given preferential treatment (to the possible exhaustion of bandwidth for BE traffic); to configuring a proportion of resources (e.g. output bandwidth) at each Layer-3 hop for each DSCP or group of DSCPs; to more full-blown metering (measurement), policing (distinct handling of out-of-profile packets), and output link resource allocation

(bandwidth) for each DSCP or group of DSCPs. Note this is also a minimalist class in that individual flows are not recognized anywhere in the network, not even at edges.

4. Host DS-byte marking, no signaling, some flow-recognition near edge: an extension of item 3 above, this method adds the feature that some form of flow recognition occurs near the edge. Thus manually configured resource commitments might be made, not only to particular DSCPs, but also to particular "flows". Here the "flow" might be characterized by destination prefix (e.g. 10mbps towards A.B.C.x), source prefix, or even full 5-tuple. The idea is that once a packet is analyzed and handled (at some level of granularity) at the edge, the packet is subsequently only treated as part of a larger aggregate. Also note that if 5-tuples are recognized by the first Layer-3 device, then one could arrange for the Layer-3 device to mark the DS-byte, rather than requiring the host to do it.

5. Local signaling, static inter-AS provisioning: This introduces the concept that a host or application might dynamically signal for resources. Also, a bandwidth broker and policy server might apply administrative policy as to which applications are allowed to emit flows that receive preferential treatment, and dynamically keep track of intra-AS commitments. Layer-3 devices might be reconfigured by the BB as new resource commitments are made. In the simplest form, the links across AS boundaries are still statically provisioned. Note that this requires careful monitoring of links to destination ASes. Two source ASes (AS-1, AS-2) might submit packets in-profile for their individual agreements with a transit AS (AS-t), but if both packet streams were destined for a third AS (AS-3), any bandwidth committed towards the destination AS might be easily exceeded at the output link from AS-t to AS-3. Protocols under discussion for the intra-AS signaling include adaptations of DIAMETER and RSVP. The method most often discussed for a BB to control intra-AS Level-3 devices is COPS, although several home-grown methods where a BB telnets to a Level-3 device to configure it also exist.

6. Single-ended signaling, with inter-BB communication: This extends item 5 by keeping the notion that a host or application might express needs to an intra-domain BB, and adding the notion that BBs in different ASes communicate with each other. The inter-AS communication allows for dynamic adjustment of the commitments made across the boundary between AS-1 and AS-t. Note that the agreement between AS-1 and AS-t is pairwise, but that acceptance of a new allocation level across that link may require AS-t to do some resource re-allocation internally. If the inter-AS BB communication also introduces the notion of certain resource allocation across the {AS-1, AS-t} link, but with additional information that the increase is to accommodate extra traffic towards AS-3, then AS-t has information that can propagate towards and adjust the resource allocation across the {AS-t, AS-3} link. On one hand, this extra information would lead towards more effective allocation of resources, and increase the chance that a packet will actually get preferential treatment end-to-end. On the other hand, this extra information (destination AS or prefix) is also likely to lead to an increased level of signaling activity in all affected networks. The anticipated relationship among resource allocation quantums, frequency of update, granularity of control, and certainty of commitment are still topics of research. And that's for the unicast case. The multicast case is even more "interesting."

7. Double-ended signaling, inter-BB communication: In [IntDiff], a 3-part mechanism is proposed. RSVP is used to signal resource requirements in the source AS-1. Such RSVP messages are tunneled through intermediate ASes (AS-t), without elements in AS-t acting on them directly. The RSVP messages, upon arrival at AS-3, are used in AS-3 for intra-AS (AS-3) resource allocation.

8. Full RSVP end-to-end: This is not part of DiffServ (actually, it derives from IntServ). But it is presented as the other "pole" method for comparison. Resource allocations are signaled via RSVP, and some amount of state is installed to keep track of and act on commitments for each flow in ASes along the entire path {AS-1, AS-t, AS-3}. It is the existence of signaling for each flow, the establishment of per-flow state in the transit networks, and the need for maintenance of this state that have led many to speculate that full-blown RSVP is likely to not be scalable in very large cores (e.g. AS-t), and led to the efforts underway in DiffServ.

There are many directions that the DiffServ architecture can take as it evolves to provide end-to-end reservations. In particular, there are significant trade-offs among signaling complexity, administrative simplicity, state, trust, policy expression, strength of assurance, and scalability. Within the QBone testbed, we aim to provide room for ample experimentation to explore this solution space.

The QBone Initiative

Goals

Although the evolving IETF DiffServ architectural framework offers a promising approach to overcoming the scalability, interoperability, and administrability problems that have plagued previous QoS efforts, the strength of the architecture and the mindshare momentum currently behind it do not alone guarantee success. DiffServ has not yet been evaluated in the wide-area, and the architectural framework begs many questions and leaves many difficult research, engineering, and policy problems unaddressed. For example, it is far from clear how to perform efficient admission control for connectionless networks, what implications DiffServ will have for traffic engineering, how to design protocols for interdomain DiffServ reservation setup, how to provide for advanced reservations (*e.g.* to support scheduled distance learning courses), or what protocols and admission control algorithms are needed to support multicast DiffServ.

Because of the research and higher education community's openness and need to find common solutions to enable new advanced applications, and because of the tolerance of its applications developers and users for pre-production internet services, Internet2 is uniquely situated to build the first interdomain testbed for differentiated services and to begin to tackle the problems mentioned above. In September 1998, Internet2 announced the QBone initiative with a Call for Participation (CFP). The primary goal of the CFP was to identify a small and focused initial group of participants who would cooperate to build an open and heavily instrumented testbed. In this testbed, experimental interdomain differentiated services could be deployed, debugged, analyzed, and refined by networking engineers and researchers working in close collaboration with the users and developers of new advanced networked applications.

Organization

The response to the CFP was overwhelming - 37 proposals were submitted from more than 73 organizations. Most proposals were of extremely high quality, and many came from teams already representing collaborations between multiple organizations. A subcommittee of the Internet2 QoS Working Group reviewed the submitted proposals carefully with the primary goal of identifying a small initial group that was topologically contiguous and able to participate in building the initial *inter*domain testbed. This group was dubbed the QBone Interoperability Group (QIG).

The working group also announced the formation of the QBone Solutions Group (QSG), a second prong of the QBone initiative. The focus of the QSG is to be on supporting research and engineering relating to the deployment of *intra*domain differentiated services. This group will participate in a broad range of discussions on engineering and deployment issues, and will include both teams that plan to join the QIG and teams that do not anticipate joining this core group but that are interested in working together to share DiffServ implementation experiences and find common solutions. Participation in this group is open to the entire Internet2/NGI community. The major initial activity of this group will be the planning of a large information sharing and problem solving workshop to be held in the spring of 1999.

Participation

Current participants in the QIG include vBNS, Abilene, ESNet, NREN, CA*Net2, SURFnet, TransPac, MREN, NYSERNET, NCNI, and the Texas GigaPoP, as well as numerous universities and labs. A map showing the set of initial participants and their connectivity is shown in Figure 2. A full listing of participants with links to individual project pages may be found from the QBone Home Page [QBone].

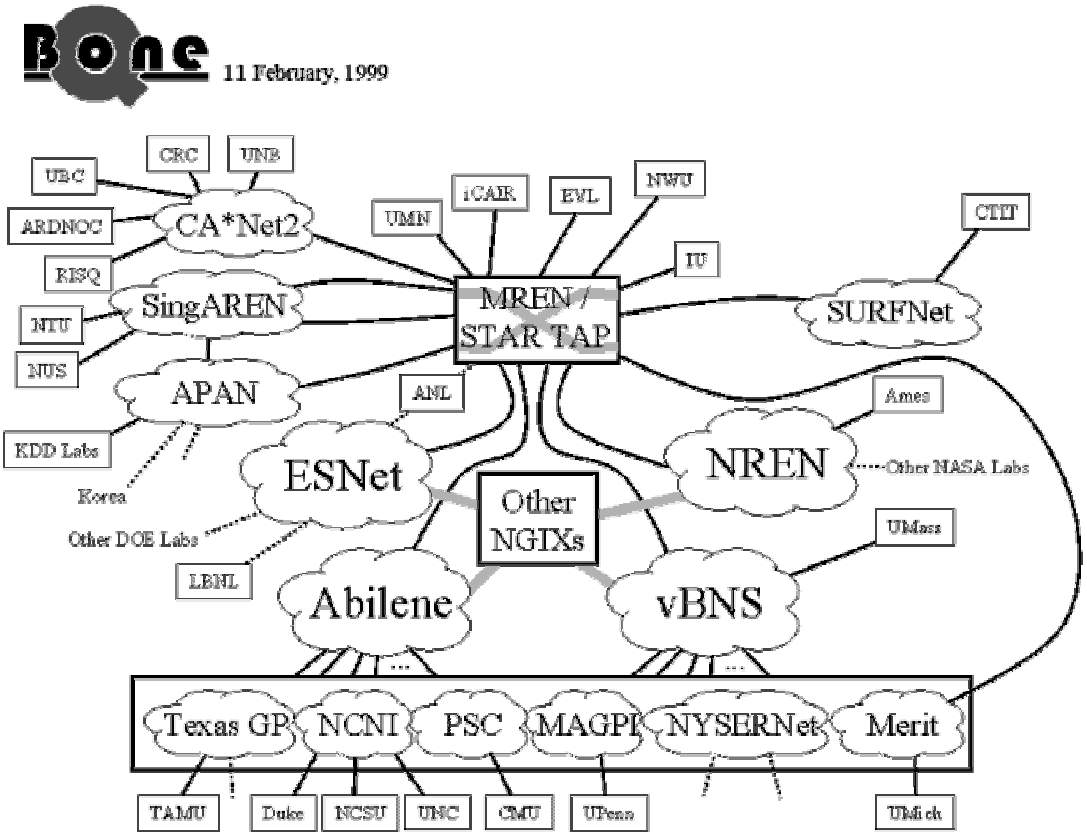

Figure 2. Initial QBone Participants and Connectivity. Actual connectivity and participant groups will change as deployment progresses.

QBone Architecture

The QBone architecture seeks to remain consistent with the emerging IETF standards for DiffServ. In addition to specifying which subset of the IETF DiffServ architecture [RFC2475] must be implemented, the QBone Architecture Draft [QBoneArch] specifies a QBone Premium Service, consequent minimum requirements for an interdomain SLA, requirements for an integrated measurement infrastructure, and a set of common operational practices for establishing interdomain reservations.

QBone Premium Service

The QBone Premium Service will make interdomain, peak-limited bandwidth assurances with virtually no loss and virtually no delay or jitter due to queuing effects. QBone Premium Service (QPS) exploits the Expedited Forwarding (EF) per-hop forwarding behavior, which is specified in [EF]. As QBone deployment progresses, the QBone Premium Service will increasingly come to resemble the virtual leased line Premium service proposed by Van Jacobson and initially demonstrated across ESNet to the show floor of SuperComputing '97 [SC97].

A QPS reservation is for a specified peak rate of EF traffic and a specified "service MTU". It offers the following transmission assurances:

- **Low loss**
  This should be very close to zero, but will not be quantified in this service definition.

- **Low latency**
  Queuing delay will be minimized, but no assumptions regarding minimal latency routing are made.

- **Low jitter**
  Delay variation due to queuing effects should be no greater than the packet transmission time of a service MTU sized packet at the subscribed rate; no assumptions about jitter due to other effects (e.g. route instability) will be made.

Minimum Requirements for QBone SLA

Consistent with the DiffServ architectural model, all service level agreements (SLAs) are determined bilaterally between adjacent QBone networks (dubbed "DS Domains" in [RFC2475]). However, to implement the QBone Premium Service, certain minimum requirements for any QBone SLA must be met. The following is a list of recommendations ("shoulds") and requirements ("musts") for any QBone SLA supporting the QBone Premium Service. The list assumes a bilateral SLA between an upstream QBone DS domain **U** and a downstream QBone DS domain **D**.)

- Within the QBone, the DS-byte Codepoint 101110 should be used for the EF PHB.

- **D** must respond to reservation requests from **U**. The protocol by which a reservation is established specifies how **D** must respond to admission requests.

- A necessary part of any SLA is a Traffic Conditioning Agreement (TCA) that specifies how traffic is conditioned and policed on ingress. The TCA is a dynamic component of the SLA, which may need to be adjusted with the creation or tear-down of every reservation across the demark. To implement QPS, a TCA must specify:
  a) Traffic conditioning
  First, ingressing traffic must be conditioned into EF and non-EF traffic. Then EF traffic may be conditioned into either a single EF Behavior Aggregate (BA) or a set of EF behavior aggregates, each of which could be defined by destination prefix or by the egress link in domain **D**.
  b) Traffic profiles
  A traffic profile must be specified for each behavior aggregate. Given a peak rate **R** and "service MTU" **M**, the traffic profile is defined by a token bucket with a token rate of **R** bytes per second and a bucket depth of **M** bytes.
  c) Disposition of Excess Traffic
  Traffic within a BA that exceeds the aggregate's profile should be discarded.
  d) Shaping
  Shaping of individual traffic flows or aggregates may be supported by ingress/egress QBone boundary nodes as an option.

- Ingressing EF traffic conforming to the traffic profiles of the TCA will be given EF treatment across DS domain **U** toward its destination. The EF PHB requires the same low

loss, low latency, low jitter packet delivery assurances discussed for the QBone Premium Service above.

- EF packets should be routed identically to packets with the Default PHB (best-effort).
- Every SLA must specify the jitter assurance made to conforming EF traffic.

Integrated Measurement Infrastructure

An integrated measurement infrastructure is key to understanding and debugging end-to-end QoS performance. The QBone Architecture requires that a set of performance parameters be collected at the ingresses and egresses of each participating QBone DS domain. These data are to be collected through both active and passive monitoring and are to include such parameters as:

- One-way packet loss
- One-way packet delay
- One-way packet delay variation
- EF load (bandwidth of a link currently devoted to EF traffic)
- EF load variation
- EF load vs. EF commitments (contracted profiles)
- Reservation load
- Reservation distribution
- Application-specific performance metrics

Bandwidth Broker

To allow QBone deployment and experimentation to begin as soon as possible, reservations will initially be long-lived and will be established manually, relying on human operators to make admission control decisions, provision appropriately and configure edge devices. This manual method of reservation will adhere to a set of common operational practices agreed upon by QIG participants. It is expected that the complexities of the manual resource allocation, device configuration, and policy management will soon overwhelm the capabilities of a human operator.

To address the overload of the human operator, it has been suggested that integral to the DiffServ architecture should be a "bandwidth broker" - an automated admission control agent that makes resource management and policy decisions in response to requests for bandwidth reservations. Within the Qbone initiative, a Qbone Bandwidth Broker Advisory Council (QBBAC) has been formed to recommend bandwidth broker solutions and to develop a pre-standards inter-domain bandwidth broker signaling protocol for experimental deployment in the QBone. This group is being led by Susan Hares of Merit Network. The focus of the group is to "initiate the exchange of ideas, among the Advisory Council and to establish some agreements on bandwidth broker issues for preliminary implementation and interoperability testing." [BBReq]

The first challenge in forming these recommendations and specifications is to precisely define what a bandwidth broker is. Due to the variety of approaches used by researchers and commercial organizations, there is a wide scope of what has been called a "bandwidth broker." The task of the QBBAC group is to gather the best ideas from these varied approaches and encourage their use in the QBone initiative.

With so many good ideas and so little real experience with bandwidth brokers, the QBBAC has taken the approach of encouraging the research and development of many different approaches to the bandwidth broker for intra-domain use within a network. Just as a gardener allows many wild flowers to grow and blossom in his patch of ground, the QBBAC decided to encourage different approaches to bloom and grow into bandwidth brokers. From these recommended intra-domain bandwidth brokers, each DiffServ domain will be able to select one or more for early experimentation in the networks in the QBone initiative. As networks experiment with the different approaches, network operators can evaluate how useful each approach is.

The QBBAC is, therefore, focusing its efforts on two areas: (1) clarification of bandwidth broker terminiology and the role of BB in the Internet2 QBone architecture; (2) the development of mechanism to support bandwidth reservations from one DS domain to the next. To share bandwidth between two domains, the two domains must agree on an inter-domain BB signaling protocol. The QBone BB Advisory Council has begun to define a pre-standards BB-to-BB inter-domain signaling protocol. Inter-Domain bandwidth broker implementations will use this common inter-domain signaling protocol.

What is a Bandwidth Broker?

A bandwidth broker manages the QoS network resources within a given DiffServ domain based on the policy set by the Service Level Agreements (SLAs) that the service provider makes with its users/clients, and adjacent networks that provide it connectivity to other parts of the Internet.

An example of a Service Level Agreement (SLA) in an academic environment is an agreement between the campus network and the High Energy Physics department. In this agreement, the High Energy Physics department needs 100MB of high priority traffic through the network to run a joint experiment with a remote national laboratory. The Physics department agrees to pay the University networking group for the 100 MB premium bandwidth from the edge of the campus network to Harvard. The Biology department has a joint experiment that needs needs 50 MB of premium bandwidth to a remote biology laboratory. The University network contracts for 200MB of premium bandwidth through the GigaPoP. The GigaPoP has 4 campuses requesting 200 MB of premium service, and it negotiates with the vBNS or Abilene for 800 MB of traffic to be passed through Abilene. Each of these service level agreements expresses the business agreements of each network so that the network infrastructure can be provisioned to meet users' needs. These service level agreements also imply admission control decisions on what can enter the network.

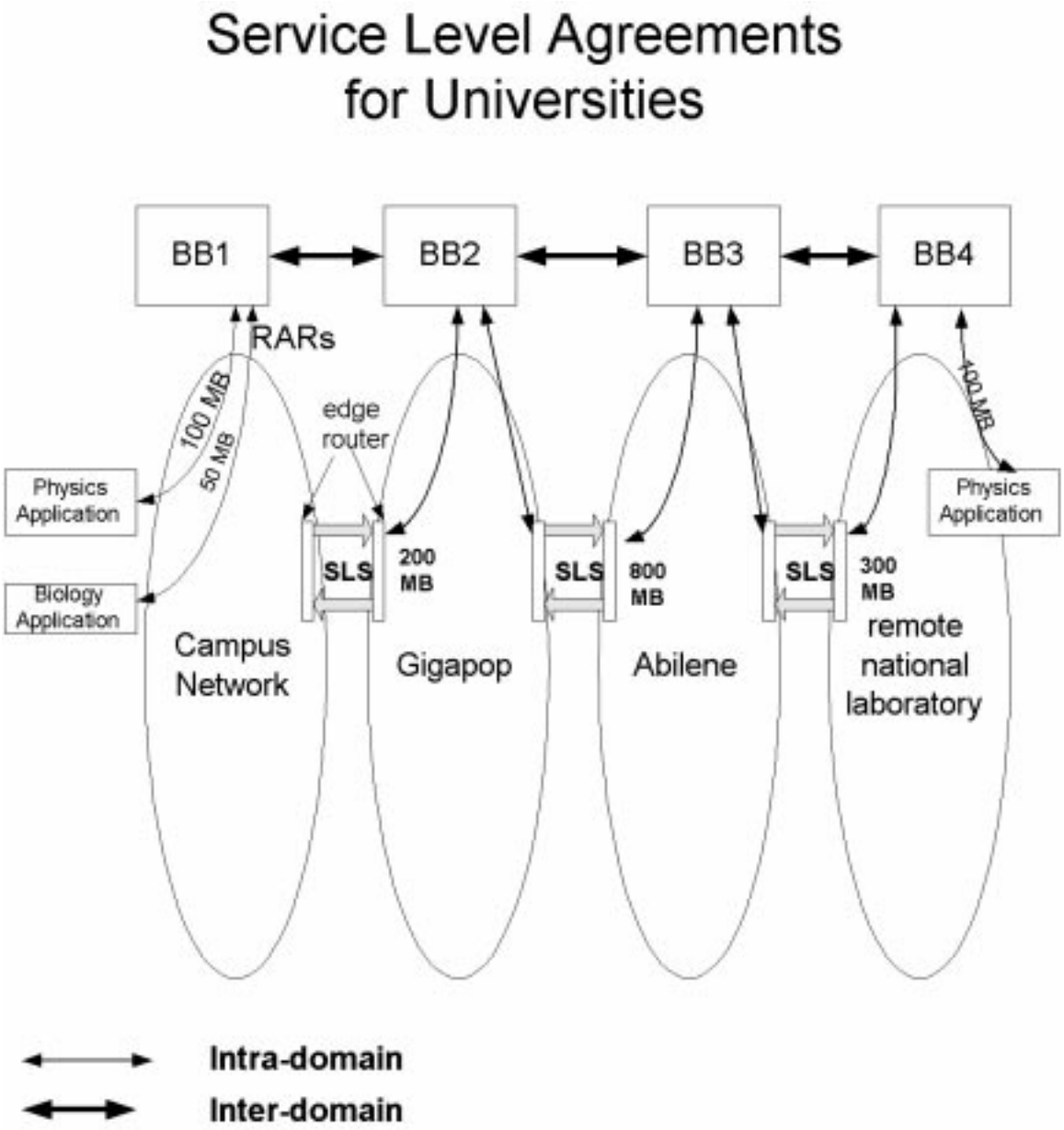

Figure 3 - Service Level Agreements on Bandwidth for Academic institutions

The connection admission control(CAC)decisions are based on policy for use premium bandwidth within a network and between two networks. For example, if the same High Energy Physics department has an experiment that once per week needs to send 150 MB of premium traffic, the physics department needs to add this fact to its SLA. The campus network translates this SLA addition into a policy statement that allows one period of 150 MB traffic per week. The Physics department's leaf router is configured to enforce the campus network policy that the premium traffic should normally be 100 MB but a 150 MB request can be made once per week. The request for 150MB is initiated by the Physics application, which sends a request for this additional bandwidth to the bandwidth broker. The Bandwidth Broker configures the leaf router at the Physics department to allow the 150 MB of data for the premium QoS services. If a second request for 150 MB occurs during a week, the BB will not change the leaf router to allow 150

MB of data to pass through the network. The admission control for the premium packets entering the campus network depends on the partnership between the bandwidth broker's policy decisions and the router's mechanism to enforce admission control for premium service.

In the IETF, the policy- and QoS-related working groups such as Policy Framework, RAP, DiffServ, and AAA have begun working to denote this router at the trust boundaries as the Policy Enforcement Point (PEP) and the Bandwidth Broker as the Policy Decision Point (PDP). The collection and recall of policy within a bandwidth broker is like the descriptions of the policy servers from the Policy Framework working group descriptions.

As the example in the academic network demonstrates, the bandwidth broker contains a means to keep track of the bandwidth resources, the policy to determine what steps to take if the bandwidth does not meet the SLAs, and a means of communicating the right information to the PEPs. The bandwidth allocation policy is just a portion of the full policies of a network. A group of policy manager devices within a network will interact to enforce the policy of the network on Connection Admission Control or usage. Given this background from the IETF, the QBone Bandwidth Broker Advisory Council has begun to define the bandwidth broker on the basis of the terms below. All quoted definitions below are taken from the Internet2 Qbone Advisory Council working document "A Discussion of Bandwidth Broker Requirements for Internet2 Qbone Deployment" (version 0.3)[BBREQ]

Bandwidth Broker (BB)

> A bandwidth broker (BB) manages network resources for IP QoS services supported in the network and used by customers of the network services. BB may considered a type of policy manager (see Policy Manager definition below) in that performs a subset of policy management functionality.

Policy Manager (PM) or Policy Server (PS)

> A policy manager (PM) or policy server (PS) typically manages the access of users to network policy services. As part of the process of admitting users to access policy services, a PM may employ a BB for CAC, as described above.

Connection Admission Control (CAC)

> Connection admission control refers to the process, performed by the BB, of admitting connection requests to the network based on available resources in the network. The determination of available resources may be done on a static or dynamic basis.

Domain

> A network domain, in general, refers to a collection of nodes (hosts, routers, etc.) and a set of links connecting them. In the context of this document, we refer to a domain as that collection of nodes and links that are under the control of the BB. Although it is not necessary, a BB's domain is usually associated with an autonomous system (AS), typically operated by a single administrator.

Inter-Domain Communication

Inter-domain communication refers to the protocol messages and control data that gets exchanged between BBs in adjacent domains.

Intra-Domain Communication

Intra-domain communication refers to the protocol messages and control data that gets exchanged between a BB and the nodes (usually edge devices) within that BB's domain.

Resource Allocation Request (RAR)

A RAR refers to a request for resources (or service) from an individual user to the BB of that user's domain. If the request is for resources for traffic to a destination(s) outside of the user's local domain, the admission control may be performed based on the SLD(s) in place with adjacent domains.

Service Level Description (SLD)

A SLD refers to the particular information relative to the BB and the network devices in order to support an SLA in that network. The SLDs are a translation of a Service Level Agreement (SLA) into a set of information that will aid automatic allocating and provision of QoS resources within network devices. The BB collects and monitors the state of QoS resources within a domain/network. This collection can be either dynamic, via an interface to the network routing information or via an interface to network configuration database. Information in the SLD is generally on the level of network ports, IP addresses, (aggregate) data flows, resources/bandwidth, etc.

Bandwidth Broker Architecture

The bandwidth broker manages QoS resources based on Service Level Descriptions (SLDs). The available QoS resources and the policy information in the SLDs are used to determine what requests for QoS resources can be honored. In the example above, the SLD of the High Energy Physics group allowed the short term use of 150MB of data only once per week. Within a domain, the BB needs to verify that QoS resources are sufficient to honor the existing SLDs.

The BB also monitors the use of QoS resources within the local domain as well as the use of inter-domain QoS resources. The BB coordinates SLDs with other domains via inter-domain communication. Across boundaries, the SLD will be an aggregation of QoS bandwidth requests within the domain of a particular QoS service type (i.e. DSCP). Within a domain, resources are allocated to applications by means of the Resource Allocation Requests (RARs). It is the responsibility of the BB to coordinate allocation and provisioning of the aggregate resources of the SLDs, into and out of its domain, with those resources requested via RARs. Another responsibility of the BB is to allow preemption of a current connection for a higher priority RAR.

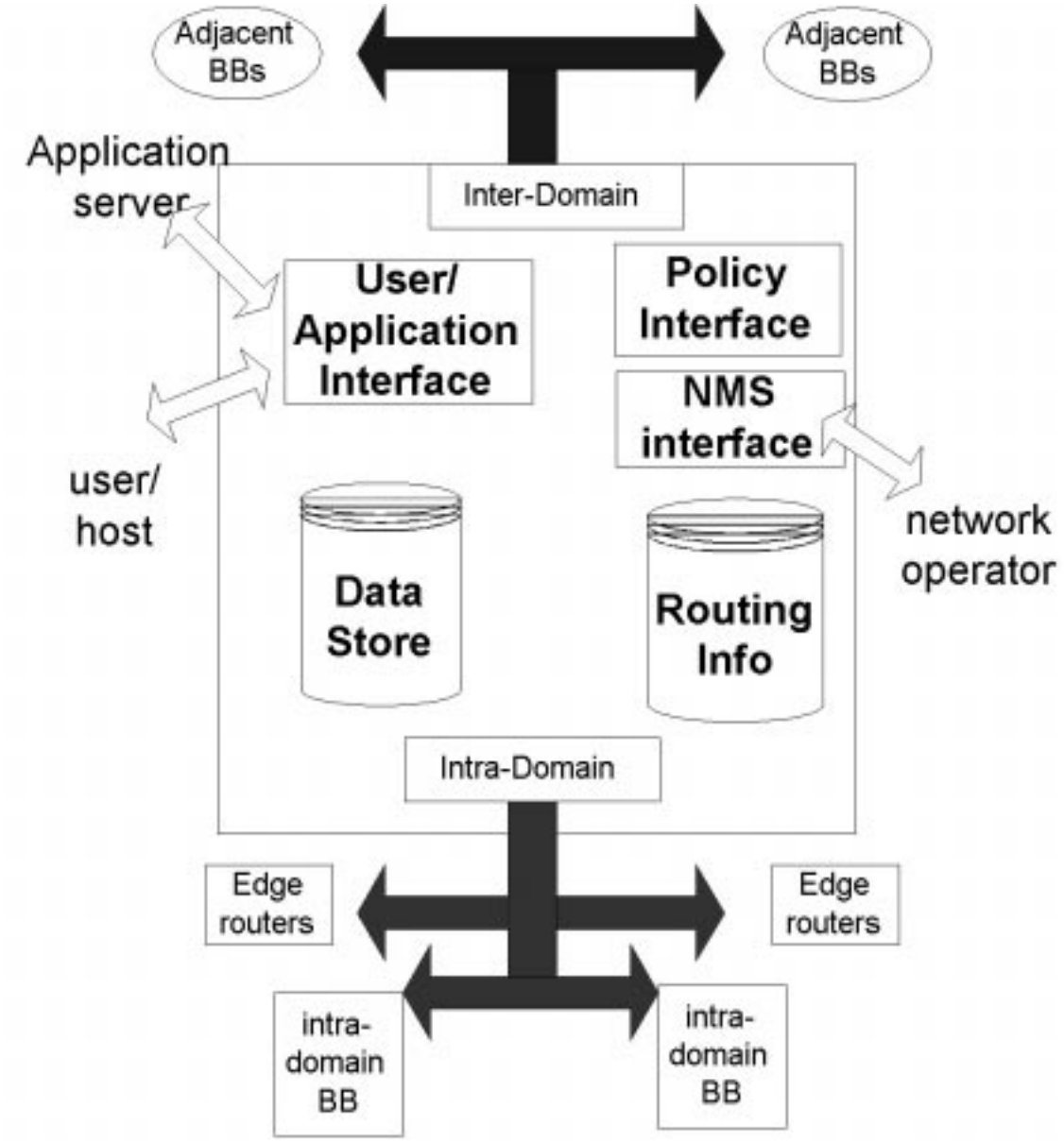

Figure 4 - Bandwidth broker Architecture

To implemement the above mentioned requirements, an architecture has been proposed. This architecture includes the components illustrated in figure 4, and described below.

The user/application interface is responsible for receiving requests from an application server, a user on a host, a network operator, or from a router within its domain. This allows an application to request the bandwidth directly or via an application server (eg. H.323).

The intra-domain interface allows the BB to reconfigure the leaf and edge routers in order to provision the QoS bandwidth. In complex configurations, multiple bandwidth brokers may be required within a single domain. In this situation, the BB must communicate with other bandwidth brokers within its domain to coordinate policy decisions and allocate bandwidth via the intra-domain interface.

The BB communicates with BBs in adjacent domains via the inter-domain interface.

The NMS Interface enables the network operator to manually configure QoS mechanisms via a GUI or a command line interface. This interface provide the network operator with the ability to adapt the network provisioning and traffic management to meet unusual or critical needs.

The routing information repository allows the bandwidth broker to store information from intra-domain (OSPF) or inter-domain (BGP) routing that pertains to the QoS provisioning. A BB implementation developed in Europe [Telia] gathers this information via a routing interface to a GateD routing daemon.

The Data Repository is used by all components of the BB and may be shared with a remote policy manager via the policy manager interface. The remote policy manger may use the Policy Manager interface to coordinate SLD and network resources between Policy Decision points such as Network Access Servers (NASes) and many bandwidth brokers to support admission control to a particular network.

Inter-Domain Bandwidth Brokers Communication Models

Figures 3 and 4 show two possible notification and response models for the inter-domain communication in early phase of deployment: the end-to-end model, and the immediate response model. BB1, BB2, and BB3 represent the BBs for AS1, AS2, and AS3 shown previously in Figure 2. The current plan of the Bandwidth Broker Advisory Council is to investigate the end-to-end communication model because it provides richer signaling experiments.

End-to-End Model

Figure 5 shows the end-to-end notification/response signaling. A user request is received by BB1 in domain AS1. After determining that local resources are available and the request will not violate the SLD with the downstream domain, BB1 notifies BB2 of the effect of the request on the aggregate resources provisioned between the domains. BB2 informs BB3 who informs BB4. If local admission control fails at any BB along the way, notifications stop flowing downstream and a (negative) response indicating where the failure occurred is immediately sent back upstream to the originating BB.

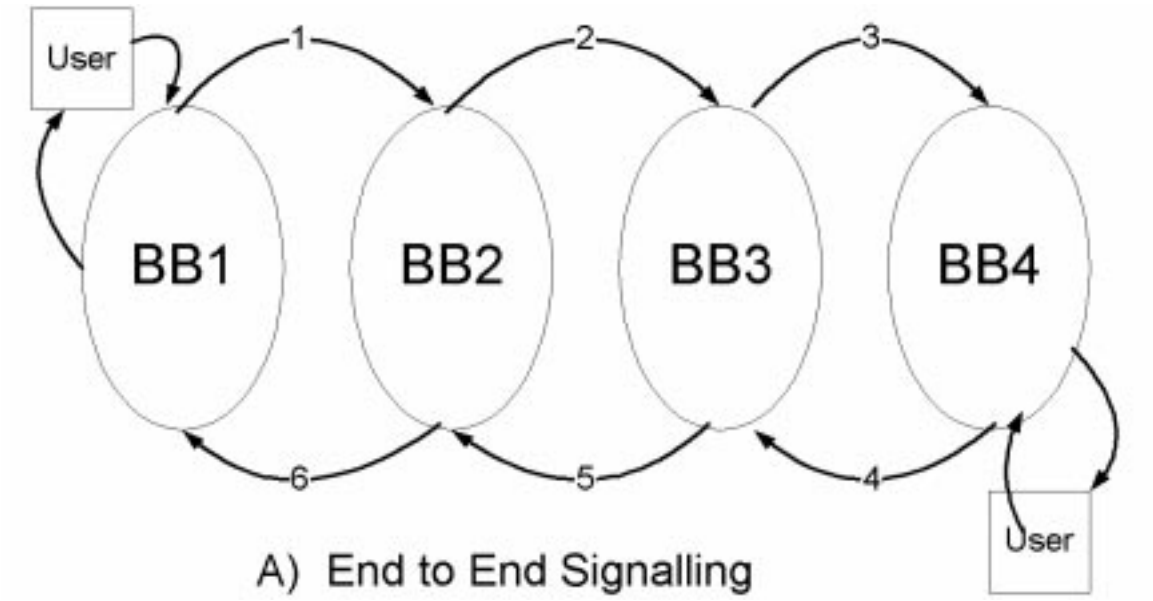

Figure 5: Inter-domain Notification/Response Examples

Immediate Response Model

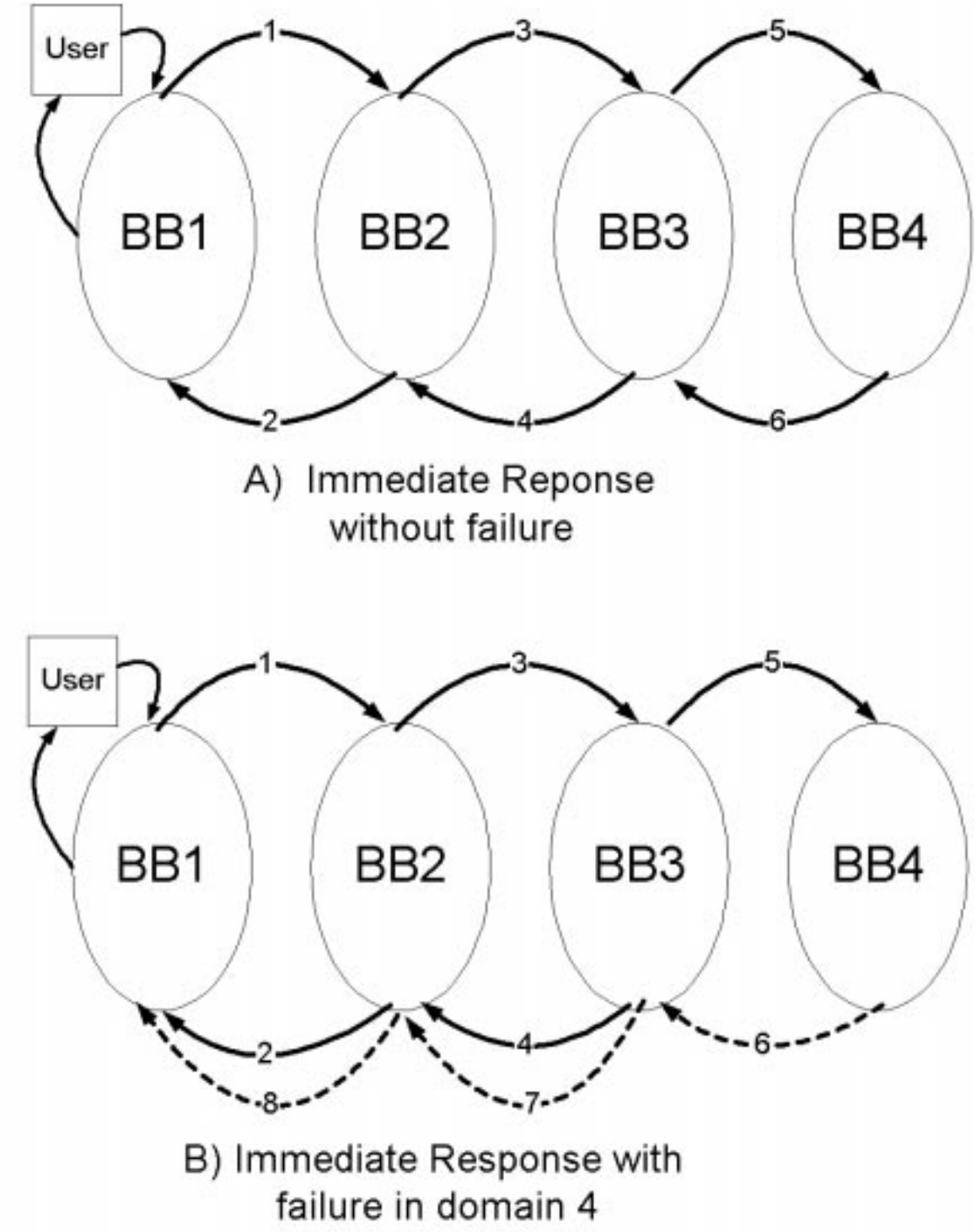

Figure 6 - Immediate Response

Figure 6 shows the immediate response model for inter-domain communication. Figure 6A shows a successful bandwidth negotiation. Figure 6B shows a failure in bandwidth allocation occurring in domain 4.

In figure 6A, BB1 performs local admission control. Admission is granted and notification is sent to BB2. BB2 grants admission and sends a positive response back to BB1. BB2 also notifies BB3. BB3 accepts the notification and a positive response is sent back to BB2. BB3 sends notification to BB4. BB4 accepts the notification and a positive response is sent back to BB3.

Figure 6B shows what happens if BB4 rejects the notification - a negative response is generated to BB3 (step 6). Upon receiving the negative response BB3 sends back a negative response to BB 2 (step 7), who in turn sends back a negative response to BB 1 (step 8). Due to this in the immediate response model, we see that BB 3 may receive data from domain 1 and 2 without being able to deliver it to domain 4.

One of the problems with the immediate response model is that a BB may accept the request from its local user upon receiving a positive response from the neighboring BB and without any indication that QoS resources for the traffic flow will be available in any other downstream ASes. Traffic may be sent through domain 1 and domain 2 only to be dropped in domain 3 due to insufficient provisioning in that network. This is not possible in the end-to-end model since notifications and responses go from BB to BB to the destination AS and back before the user's request is admitted.

Since there is so little experience with these models (or a hybrid of these models), it makes sense to allow experimentation with both of them. If a guarantee of end-to-end QoS is required before traffic is sent due to a high traffic load, the model in Figure 3 would be more appropriate. However, if an application simply wants to get the best service, as far as possible, the immediate response model of Figure 4 would suffice. To facilitate both possibilities, inter-domain BB-to-BB signaling protocol needs to indicate the type of response method requested.

Deployment of Inter-Domain Bandwidth Brokers

Within the discussion about the Inter-Domain Bandwidth Broker in the QBBAC, there is a long list of proposed capabilities for the Inter-Domain signaling between the bandwidth brokers. As the Qbone BB Advisory Council has begun to define a pre-standards bandwidth broker to bandwidth broker inter-domain signaling protocol, the group has taken the approach of specifying a subset of the capabilities, implementing that subset and deploying implementations of that subset. Because of this approach, the deployment of Bandwidth Broker technology will be deployed in incremental steps of added technology, called phases, described below.

The first three phases use static Service Level Descriptions (SLDs), instead of dynamic re-negotiation. Dynamic re-negotiation of Service Level Agreements via dynamic SLDs allows the adjustment of resource commitments between two domains. If resources are adjusted, the internal network devices must be reconfigured to match the adjustment. These three phases limit the BB communication to simple notification and response without requiring the adjustment of resources on internal switches and routers.

Phase 0 - Local Admission

> Phase 0 is referred to as local admission because from the BB point of view, requests for resources are admitted into the network with admission control relative only to the local domain where the request is received. Phase 0 will have automatic resource allocation at the

intra-domain level, but use "static" manual provisioning at the inter-domain level. If the request is accepted in the local domain it is admitted into the network without further checking if resources will be available all the way to the actual destination. In most early cases, the local domain management will use simple mechansims such as allocating only 10 premium services users at a time. After the 10^{th} user is added, no more premium service is allocated until a user gives up the premium service "token" and a new user can be added as the new 10^{th} user.

Phase 1 - Informed Admission

Phase 1 is referred to as informed admission because the local BB makes admission control decisions based on information from downstream BBs. The bandwidth brokers will communicate in a peer-wise fashion. The information passed to a BB from its peer can originate from the peer or from a remote peer. The inter-domain signalling can either be an End-To-End signalling (figure 5) or a immediate response signalling (figure 6).

At each level of inter-domain communication, the resource requests will be aggregated. The communication between BBs will be based on "static" SLAs. The BB to BB communication provides a BB with information about destination of the aggregate traffic, and the necessary resources to support that traffic. If the "static" SLAs are not sufficient, the additional needs may be communicated via the network management interface.

Phase 2 - Guaranteed Admission

Phase 2 is referred to as guaranteed admission because it guarantees that if all BBs from a source to a destination agree that these are sufficient resources, the packets will not get dropped due to a transient data bursts. The "admission" comes from the phase 1 "informed admission" control where the local BB makes admission control decisions based on the information from downstream BBs. As in phase 1, if the "static" SLAs are not sufficient, the additional needs may be noted to the network management interface inside the bandwidth broker.

Phase 3 - Dynamic SLD Admission

In phase 3, the BBs in addition to supporting inter-domain signalling will be able to dynamically set up the new SLDs. One or more SLDs can be associated with an SLA. A bandwidth broker will support for configuring traffic conditions are the edge routers at the domain boundary. Additionally bandwidth brokers may have the ability to configure routers interior to their network.

Phase M - Multicast Bandwidth Negotiation

Multicast bandwidth negotiation will be investigated after the unicast bandwidth broker issues are resolved.

Security Considerations

The primary aim of the DiffServ architecture is to provide different levels of service to different traffic streams on a common network infrastructure. Any techniques used to implement such

resource reservations will cause some traffic flows to receive better treatment than others. Two methods of creating a denial-of-service attack are altering the DiffServ field or by injecting packets with the DiffServ field set to codepoints that make the packets receive enhanced service levels. This theft of premium resources could result in a denial-of-service attack when the modified or injected traffic depletes the resources available to forward it and other traffic streams.

In a DiffServ domain, any client that wishes to establish a communication channel with a set of guaranteed resources makes a request of a reservation manager. For our discussion, limiting the managed resource to bandwidth only, the resource reservation manager is called a bandwidth broker. A bandwidth broker should be able to provide the requesting client with a reservation for a secure connection either within a single domain or across domain boundaries, for an end-to-end reservation.

Intra-domain reservations

When an intra-domain BB receives a reservation request, its first actions are to determine and verify the identity of the requesting client. Depending on the environment in which the BB operates, it may or may not be able to use the environment's authentication mechanism to carry out these tasks. For example, the Globus environment [V1] proposes to use the Globus [V2] authentication mechanism. All those who exercise authority over the allocation of bandwidth can impose restrictions on its use. Such restrictions, which may constitute the policy governing access to the resource (i.e., the link), could be based on time of day, source address or address prefix, group memberships or application traffic type, or any other form of access control [V3].

Once the requesting client has been authenticated, the BB must use a previously agreed upon authentication mechanism to determine that all access control policy checks have been satisfied. At this point a successfully established reservation could be represented by a token or an encrypted certificate. All corresponding domain policy checks will entail the BB having access to a certification authority.

Inter-domain BB and end-to-end reservations

Security within the inter-domain BBs, like security within other inter-domain protocols such as BGP, has three components: peer identity, link, and data. The BBs in different domains need to establish a bilateral peering (trust) relationship between remote peers. The BB can identify and validate its neighbor by means of authentication tokens, certificates, or pre-configuration. Security of data going across a specific link can be done by using IPSec or other mechanisms which guarantee security across a single link. An end to end secure path is set up by establishing secure bilateral peer relationships among the BBs from the source domain to the destination domain.

Data security allows the originator of the RAR, to secure an individual request. Normally, RARs will be aggregated by the local Bandwidth Broker. Figure 1 showed the aggregation of the RARs into a single BB to BB exchange. Once the Bandwidth Broker aggregates the RARs, the combined aggregate sent as an SLD will need to be secured as a new bandwidth request via certificates or authentication tokens. In a few cases, it may be desireable to allow SLD from the originating domain to be passed intact to the remote domain. For example, if in figure 1, the

aggregation of the two RARs into 1 SLD could be digitally signed and passed through from domain 1 to domain 3.

Inter-domain issues for Ingress Routers

An ingress router in a domain is always the first line of defense against any kind of service attacks based on modified codepoints. A node in a DiffServ domain that is the source of traffic acts as a ingress node for that traffic in the domain, and therefore must ensure that all traffic carries acceptable DiffServ codepoints. An ingress router may be required to modify the codepoints of incoming traffic based on previously agreed-upon service level descriptions. It becomes the responsibility of this ingress node to ensure that incoming packets are in-profile according to the codepoints, and to discard them if they are not. The ingress node may also be required to do traffic conditioning. In addition, the ingress node may need to apply authentication mechanisms to validate some incoming traffic flows, but leave others untouched if the traffic is known to be originating from a trusted source (site) or if the inbound link itself is trusted.

Interaction of Non-DiffServ to DiffServ Domains

Any links outside the purview of the DiffServ domains and/or the DiffServ network may be subject to local security policies. To ensure link integrity, security on these links may be implemented via physical control devices or by other means such as IPSec. With respect to the use of IPsec within DiffServ domain boundaries, it is worthwhile to note that the IPsec protocol currently requires that the inner header's DiffServ field not be changed by IPsec decapsulation processing at a tunnel egress node. This ensures that an adversary's modifications to the DiffServ field cannot be used to launch theft- or denial-of-service attacks across an IPsec tunnel endpoint, as any such modifications will be discarded at that endpoint. Thus defense against such attacks could consist of a combination of traffic conditioning at DiffServ boundary nodes and the security and integrity of the overall network infrastructure itself.

Conclusions

The QBone will be the first wide area test of the evolving differentiated services architecture and the first experimental deployment of interdomain differentiated services. It is envisioned that the QBone will grow incrementally as new QoS services mature. By building a highly instrumented testbed that is open and accessible to researchers and advanced development efforts, the QBone initiative seeks to advance the state of DiffServ technology. Further, by working together with the broader Internet2 community to come to terms with the profound administrative, economic, and policy implications of QoS, the QBone aims to start a process that will open the horizon for new advanced networked applications to flourish.

Acknowledgements

The authors gratefully acknowledge Guy Almes for clear-sighted guidance and moral encouragement. Additionally, we would like to recognize the invaluable editorial assistance of Ben Chinowsky on this paper.

References

[QBBAC]

QBone Bandwidth Broker Advisory Council Home Page, http://www.merit.edu/working.groups/i2-qbone-bb/

[BBREQ]

Rob Neilson, Jeff Wheeler, Francis Reichmeyer, Susan Hares, "A Discussion of Bandwidth Broker Requirements for Internet2 Qbone Deployment", version 0.3, http://www.merit.edu/i2-qbone-bb/doc/BB_Requirements_v3.doc

[DSFRAME]

*A Framework for Differentiated Services,*Y. Bernet, J. Binder, S. Blake, M. Carlson, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, W. Weiss, Internet Draft, October 1998.

[EF]

Expedited Fowarding Per Hop Behavior, V. Jacobson, Internet Draft, November 1998.

[IntDiff]

A Framework for the Use of RSVP With Diff-serv Networks, Bernet, Y., Yavatkar, R., Ford, P., Baker, F., Zhang, L., Nichols, K., Speer, M., Internet Draft, June, 1998

[I2QoS98]

Report from the First Internet2 Joint Applications/Engineering QoS Workshop, http://www.internet2.edu/qos/may98Workshop/9805-Proceedings.pdf, May 1998.

[NANO]

The nanoManipulator Home Page, http://www.cs.unc.edu/Research/nano

[QBone]

QBone Home Page, http://www.internet2.edu/qos/qbone/

[QBoneArch]

Draft QBone Architecture, http://www.internet2.edu/qos/wg/papers/draft-i2-qbone-arch-03.html, January, 1999.

[RFC2474]

Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, K. Nichols, S. Blake, F. Baker, D. Black, IETF Standards Track RFC 2474, December 1998.

[RFC2475]

An Architecture for Differentiated Services, D. Black, S. Blake, M. Carlson, E. Davies, Z. Wang, and W. Weiss, IETF Informational RFC 2475, December 1998.

[SC97]

Experience with a Class Based Queuing Demonstration, R. Nitzan, http://www.es.net/nesg/esnet-qbone-participation/cbq_test_paper.html, February, 1998.

[Quantum]

Quantum Project Home Page, http://www.dante.net/quantum/

[V1]

Ian Foster, Carl Kesselman. "Globus: A Metacomputing Infrastructure Toolkit." Intl J. Supercomputer Applications, 11(2):115-128, 1997.

[V2]

Ian Foster, Carl Kesselman, Gene Tsudik, Steven Tuecke. "A Security Architecture for Computational Grids." Proc. 5th ACM Conference on Computer and Communications Security Conference, pp. 83-92, 1998.

[V3]

Gary Hoo, William Johnston, Ian Foster and Alain Roy. "QoS as Middleware: Bandwidth Brokering System Design." Paper submitted for the Eighth International Symposium on High Performance Distributed Computing.

[telia]

"http://www.cdt.luth.se/~olov/publications", Olov Schelén, Stephen Pink: Resource Reservation Agents in the Internet. Position paper. In proceedings of 8th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'98), Cambridge, United Kingdom, July 1998. compressed postscript (58K), postscript (252K), PDF (129K)